/*QUESTION 1: The intercept from such a regression of residuals on residuals will always be zero (disregarding rounding error). Why? (Hint: what is the sample mean of the residuals from a linear regression model? And what is the formula for the intercept in a simple linear regression model?)

$Y = X\beta + e$, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & X_1 & X_2 & \cdots & X_n \\ 1 & X_1 & & & \vdots \\ & \vdots & & & \\ 1 & X_1 & X_2 & \cdots & X_n \end{bmatrix} \in \mathbb{R}^{n \times (n+1)}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$, $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

so $\hat{Y} = X\hat{\beta}$ and $\hat{\beta} = (X^T X^{-1}) X^T Y$, so $\bar{e} = \frac{1}{n} 1^T (Y - X(X^T X)^{-1} X^T Y)$

So $1^T = (Xa)^T$ where $a = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n+1}$, then $\bar{e} = \frac{1}{n} a^T X^T (Y - X(X^T X)^{-1} X^T Y)$

$\bar{e} = \frac{1}{n} a^T (X^T Y - X^T Y) = 0$, so sample mean of residuals is always zero.

$\hat{\beta}_0 = \bar{e}_{y|x_1,x_2} - \hat{\beta}_3 \bar{e}_{x_3|x_1,x_2}$, where y is fev1, x_3 is weight, x_2 is height, x_1 is age, where $\hat{\beta}_3$ is the partial regression coefficient associated with x_3 in the model, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. So $\bar{e}_{y|x_1,x_2} = \bar{e}_{x_3|x_1,x_2} = 0$

So the intercept for such a regression is zero.

```
/* QUESTION 2: How do we interpret the parameter estimates for the coefficients for regnc, regs, regw?
How do we interpret the intercept in this model?

Answer:
Interpretation of coefficient for regnc, -0.46696:
The estimated mean risk for region 2 (North Central) is 0.46696 lower than that for region 1 (North East).

Interpretation of coefficient for regs, -0.93369:
The estimated mean risk for region 3 (South) is 0.93369 lower than that for region 1.

Interpretation of coefficient for regw, -0.47946:
The estimated mean risk for region 4 (West) is 0.47946 lower than that for region 1.

Interpretation of intercept, 4.86071:
The estimated mean risk for region 1 is 4.86071.
*/
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.86071	0.24778	19.62	<.0001
regnc	1	-0.46696	0.33929	-1.38	0.1716
regs	1	-0.93369	0.32842	-2.84	0.0053
regw	1	-0.47946	0.41090	-1.17	0.2458

/*QUESTION 3: What region did we make the reference region by using the above code?

Answer: We make region 1 (North East) as the reference region by using the above code.
And the results are as follows:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.99694	4.66565	2.71	0.0484
Error	109	187.38288	1.71911		
Corrected Total	112	201.37982			

Root MSE	1.31115	R-Square	0.0695
Dependent Mean	4.35487	Adj R-Sq	0.0439
Coeff Var	30.10765		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.86071	0.24778	19.62	<.0001
regnc	1	-0.46696	0.33929	-1.38	0.1716
regs	1	-0.93369	0.32842	-2.84	0.0053
regw	1	-0.47946	0.41090	-1.17	0.2458

Write code to fit a model using a different region as the reference group
and run a regression model. Compare the output, especially the parameter estimates.
Did R^2 change? Did the ANOVA table change?

Answer: Next, we use region 2 as the reference group.
*/

```
data senic; set senic;
    if region=1 then regne=1; else regne=0;
    if region=3 then regs=1; else regs=0;
    if region=4 then regw=1; else regw=0;
run;
proc reg data=senic;
    model risk = regne regs regw;
run; quit;
```

/*Answer: After using region 2 as the reference group, we get the following results: */

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.99694	4.66565	2.71	0.0484
Error	109	187.38288	1.71911		
Corrected Total	112	201.37982			

Root MSE	1.31115	R-Square	0.0695
Dependent Mean	4.35487	Adj R-Sq	0.0439
Coeff Var	30.10765		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.39375	0.23178	18.96	<.0001
regne	1	0.46696	0.33929	1.38	0.1716
regs	1	-0.46672	0.31652	-1.47	0.1432
regw	1	-0.01250	0.40146	-0.03	0.9752

/* For parameter estimates, although the coefficients for these regions change but if we interpret
these coefficients, we will find the estimated mean risk for each region is still the same as that when we make
region 1 as the reference region. Take region 3 (South) as an example. If we make region 1 as
the reference region, then the estimated mean risk for region 3 is 4.86071-0.93369=4.86071-0.46696-0.46672,
which is the same as 4.39375-0.46672 if we make region 2 as the reference region.
Also, R^2=0.0695 and ANOVA table do not change.

/*QUESTION 4: How do we interpret the p-values in the Parameter Estimates table,
in terms of testing for differences in means by region?
What means are being compared?

Answer: Using region 1 as the reference region, since each coefficient can be viewed as the differences in estimated mean risk between reference group (region1) and region 2, 3, and 4, respectively,
and the intercept can be viewed as the estimated mean risk for region 1. If p-values are less than the significance level (0.05), then we can conclude that there is significant evidence that means of risk for two regions are different.
We interpret the p-values in the Parameter Estimates table in detail as follows:

Intercept: Since the p-value of the intercept is <0.0001, then we conclude that there is significant evidence that the mean of risk for region 1 is not equal to zero.

Coefficient for regnc: Since the p-value is 0.1716, then we conclude that there is no significant evidence that the difference in means of risk between region 1 and region 2 is not equal to zero.

Coefficient for regs: Since the p-value is 0.0053, then we conclude that there is significant evidence that the difference in means of risk between region 1 and region 3 is not equal to zero.

Coefficient for regw: Since the p-value is 0.2458, then we conclude that there is no significant evidence that the difference in means of risk between region 1 and region 4 is not equal to zero.

Means of risk for region 1, 2, 3, and 4 are compared.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.86071	0.24778	19.62	<.0001
regnc	1	-0.46696	0.33929	-1.38	0.1716
regs	1	-0.93369	0.32842	-2.84	0.0053
regw	1	-0.47946	0.41090	-1.17	0.2458

/* QUESTION 5: Write out the null and alternative hypotheses for the test conducted by "test_region". Give the distribution of the test statistic under the null, the value of the test statistic and the p-value. What do you conclude?

This is a partial F test. Full model is: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_2 census + \hat{\beta}_3 regnc + \hat{\beta}_4 regs + \hat{\beta}_5 regw$

Reduced model: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_2 census$

$H_0: \beta_3 = \beta_4 = \beta_5 = 0$

H_1: At least one of $\beta_3, \beta_4, \beta_5$ not equal to zero

The test statistic follows F-distribution, $F(3, 107)$

Test test_region Results for Dependent Variable risk				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	3.04987	2.50	0.0636
Denominator	107	1.22083		

Value of test statistic: F value = 2.5

P-value = 0.0636 > 0.05

So we do not reject the null hypothesis and conclude that there is no significant evidence that at least one of $\beta_3, \beta_4, \beta_5$ (difference of means for risk between reference region (region 1) and region 2, 3, and 4) is not equal to zero.

QUESTION 6: Conduct the overall (omnibus) F test for the model risk = length census regnc regs regw. Write out the null and alternative hypotheses. Give the distribution of the test statistic under the null, the value of the test statistic and the p-value. What do you conclude?

This is an overall F test. Full model is: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_2 census + \hat{\beta}_3 regnc + \hat{\beta}_4 regs + \hat{\beta}_5 regw$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: At least one of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ not equal to zero

The test statistic follows F-distribution, $F(5, 107)$

Test test_overall Results for Dependent Variable risk				
Source	DF	Mean Square	F Value	Pr > F
Numerator	5	14.15019	11.59	<.0001
Denominator	107	1.22083		

Value of test statistic: F value = 11.59

P-value < 0.0001

So we reject the null hypothesis and conclude that there is significant evidence that at least one of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ is not equal to zero.

```
proc reg data=senic;
    model risk = length census regnc regs regw;
    test_overall: test length, census, regnc, regs, regw;
run; quit;
```

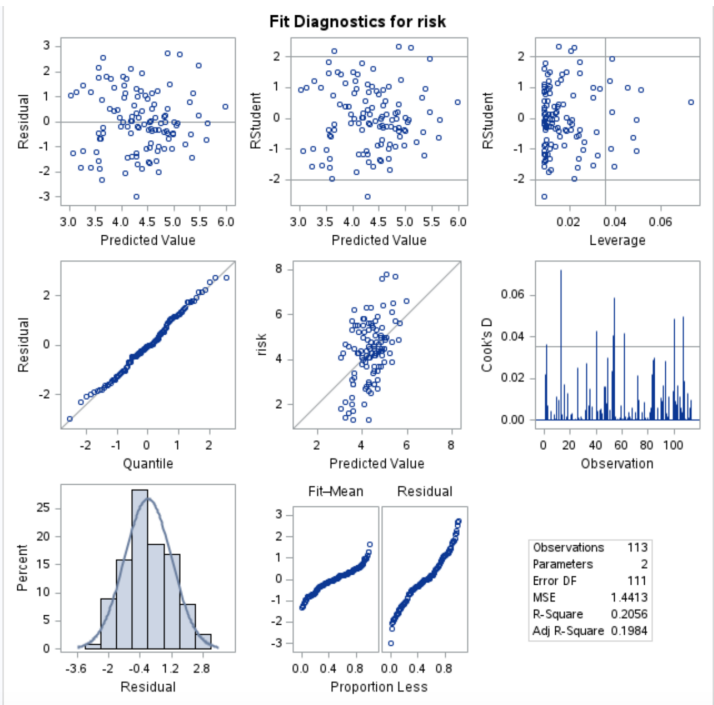
```
/* QUESTION 7: (a) A regression of risk on xray shows a highly significant relationship.
Fit the model and conduct model diagnostics (residuals analysis). Report this relationship.  Provide an
interpretation of the regression coefficients, using appropriate units.*/

proc reg data=senic plots=ResidualsBySmooth(smooth);
    model risk = xray;
run; quit;
```

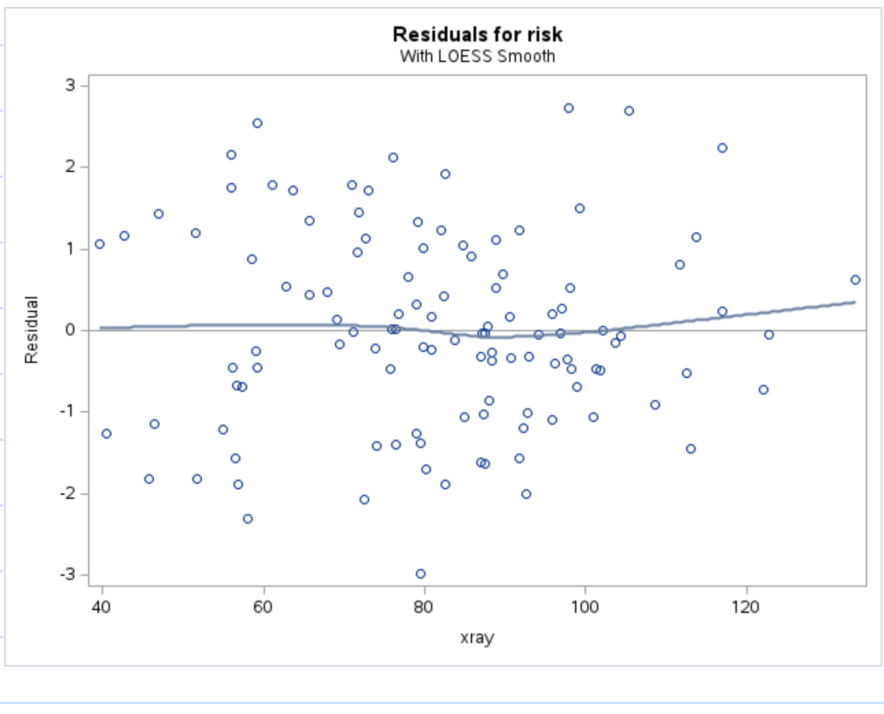
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.79202	0.49136	3.65	0.0004
xray	1	0.03140	0.00586	5.36	<.0001

Y = risk, X = xray

$\hat{\beta}_1 = 0.03140$ percents, $\hat{\beta}_0 = 1.79202$ percents

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 1.79202 + 0.03140 X_i$





From the plots above, it seems that $E(\epsilon_i) = 0$ for all i. And the error variance are roughly constant across all observations. Also, from QQ plot, all points roughly follow a straight line. So the assumptions of the normal error regression model are approximately met. So it is linear.

Interpretation for $\hat{\beta}_1$: For every one increase in the *xray*, i.e. the ratio of number of x-rays performed to number of patients without signs or symptoms of pneumonia * 100, the estimated mean infection risk will increase by 0.03140 percents.

Interpretation for $\hat{\beta}_0$: When the value of *xray* is equal to 0, the estimated mean infection risk is 1.79202 percents. However, this is only a meaningful interpretation if x=0 is reasonable.

```
/* (b) Investigators hypothesize that xray will be significantly related to risk
after controlling for beds, nurses and svcs.  What model should you fit to test this hypothesis?
Fit the model and report the results.  Do the results support the hypothesis or not? */
```

We should fit a full model: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_1 beds + \hat{\beta}_2 nurses + \hat{\beta}_3 svcs + \hat{\beta}_4 xray$

```
proc reg data=senic;
    model risk = beds nurses svcs xray / pcorr2;
run; quit;
```

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	0.86917	0.53682	1.62	0.1083	.
beds	1	-0.00033930	0.00141	-0.24	0.8106	0.00053393
nurses	1	0.00223	0.00191	1.17	0.2457	0.01246
svcs	1	0.01976	0.01162	1.70	0.0918	0.02609
xray	1	0.02858	0.00542	5.27	<.0001	0.20484

By checking t-value of xray,
Value of test statistic: t value = 5.27
P-value < 0.0001
So we conclude that xray is significantly related to risk after controlling for beds, nurses and svcs. The results support the hypothesis.

```
/*(c) Provide an interpretation of each of the regression coefficients, using appropriate units.
Also report the partial correlation of each predictor with risk.
```

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	0.86917	0.53682	1.62	0.1083	.
beds	1	-0.00033930	0.00141	-0.24	0.8106	0.00053393
nurses	1	0.00223	0.00191	1.17	0.2457	0.01246
svcs	1	0.01976	0.01162	1.70	0.0918	0.02609
xray	1	0.02858	0.00542	5.27	<.0001	0.20484

Interpretation for the intercept: When the values of *beds*, *nurses*, *svcs*, and *xray* are all equal to 0, the estimated mean infection risk is 0.86917 percents. However, this is only a meaningful interpretation if x=0 is reasonable.

For partial correlation, we only need to take square root of *Squared Partial Corr Type II*.

Interpretation for *beds*: Coefficient is -0.00033930 (percents/bed): For every one increase in the average number of beds in hospital during study period, the estimated mean infection risk will decrease by 0.00033930 percents, controlling for *nurses*, *svcs*, and *xray*.
Partial correlation is -0.02311: correlation between the residuals $e_{risk|nurses,svcs,xray}$ and $e_{beds|nurses,svcs,xray}$ is -0.02311.

Interpretation for *nurses*: Coefficient is 0.00223 (percents/nurse): For every one increase in the average number of full-time equivalent nurses during study period, the estimated mean infection risk will increase by 0.00223 percents, controlling for *beds*, *svcs*, and *xray*.
Partial correlation is 0.11162: correlation between the residuals $e_{risk|beds,svcs,xray}$ and $e_{nurses|beds,svcs,xray}$ is 0.11162.

Interpretation for *svcs*: Coefficient is 0.01976 (percents/percent): For every one increase in the available facilities and services, i.e. percent of 35 potential facilities and services that are provided by the hospital, the estimated mean infection risk will increase by 0.01976 percents, controlling for *beds*, *nurses*, and *xray*.
Partial correlation is 0.16152: correlation between the residuals $e_{risk|beds,nurses,xray}$ and $e_{svcs|beds,nurses,xray}$ is 0.16152.

Interpretation for *xray*: Coefficient is 0.02858 (percents): For every one increase in the ratio of number of x-rays performed to number of patients without signs or symptoms of pneumonia * 100, the estimated mean infection risk will increase by 0.02858 percents, controlling for *beds*, *nurses*, and *svcs*.
Partial correlation is 0.45259: correlation between the residuals $e_{risk|beds,nurses,svcs}$ and $e_{xray|beds,nurses,svcs}$ is 0.45259.

```
(d) Conduct a joint test of whether beds and nurses contribute to explaining variation in risk
after controlling for svcs and xray.
*/
proc reg data=senic;
    model risk = beds nurses svcs xray;
    test_beds: test beds, nurses;
run; quit;
```

This is a partial F test. Full model is: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_1 beds + \hat{\beta}_2 nurses + \hat{\beta}_3 svcs + \hat{\beta}_4 xray$
Reduced model: $\hat{risk} = \hat{\beta}_0 + \hat{\beta}_3 svcs + \hat{\beta}_4 xray$
$H_0: \beta_1 = \beta_2 = 0$
H_1: At least one of β_1, β_2 not equal to zero
The test statistic follows F-distribution.

Test test_beds Results for Dependent Variable risk				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1.50050	1.24	0.2924
Denominator	108	1.20629		

Value of test statistic: F value = 1.24
P-value = 0.2924 > 0.05
So we do not reject the null hypothesis and conclude that there is no significant evidence that at least one of β_1, β_2 is not equal to zero.

So beds and nurses cannot contribute to explaining variation in risk after controlling for svcs and xray.

Q8

2. Another interpretation of the coefficient of multiple determination, R^2: The value of R^2 is equal to the square of the Pearson correlation between the observed values Y_i and the predicted values $\hat{Y}_i$. In this sense, R^2 is a direct measure of how well the model fits the observed data. This can be proven mathematically, but we will forgo the proof and just show that this is true for an example:
   a. For the spirometry/FEV1 data, fit the model regressing FEV1 on age and weight and obtain the value of R^2.
   b. Obtain the predicted values for this model and the Pearson correlation between the observed FEV1 values and the predicted values. Confirm that the square of this correlation is equal to the value of R^2.

a.
```
/*Question 8*/
proc reg data=d.spirometry;
   model fev1 = age weight;
   output out = predict_data predicted=predicted_values;
run;
```

b.
```
proc corr data=predict_data pearson;
     var fev1 predicted_values;
run;
```

a.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7.14313	3.57156	141.81	<.0001
Error	68	1.71267	0.02519		
Corrected Total	70	8.85579			

Root MSE	0.15870	R-Square	0.8066
Dependent Mean	0.82972	Adj R-Sq	0.8009
Coeff Var	19.12719		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.21321	0.07887	-2.70	0.0087
AGE	Age (Months)	1	0.01051	0.00168	6.27	<.0001
WEIGHT	Body Weight (kg)	1	0.02674	0.00732	3.65	0.0005

$R^2 = 0.8066$

b.

Pearson Correlation Coefficients, N = 71
Prob > |r| under H0: Rho=0

	FEV1	predicted_values
FEV1 Forced Expiratory Volume At 1 Sec (L)	1.00000	0.89811 <.0001
predicted_values Predicted Value of FEV1	0.89811 <.0001	1.00000

Pearson correlation = 0.89811

So square of Pearson correlation

$0.89811^2 = 0.8066 = R^2$

So Pearson correlation is equal to R^2.